An Audiovisual Feedback System for Acquiring L2 Pronunciation and L2 Prosody

Grażyna Demenko[1], Agnieszka Wagner[1], Natalia Cylwik[1] and Oliver Jokisch[2]

(1) Adam Mickiewicz University, Institute of Linguistics, Department of Phonetics, Poznań, Poland
(2) TU Dresden, Laboratory of Acoustics and Speech Communication, Dresden, Germany
[1]{lin, wagner, nataliac}@amu.edu.pl, [2]oliver.jokisch@tu-dresden.de

Abstract

In recent years the application of computer software to the learning process has been acknowledged an indisputably effective tool supporting traditional teaching methods. A particular focus has been put on the application of computational techniques based on speech and language processing to second language learning. At present, a number of commercial self-study programs using speech synthesis and recognition are available. Most of them, however, focus on segmental features only. The paper presents technical and linguistic specifications for the Euronounce project [1] which aims at creating an intelligent tutoring system with multimodal feedback functions for acquiring not only foreign languages' pronunciation but also prosody. The project focuses on German as a target language for native speakers of Polish, Slovak, Czech and Russian and vice versa. The paper outlines the Euronounce feedback system and presents the Pitch Line program which can be implemented in the prosody training module of the Euronounce tutoring system.

1. Introduction

The increasing use of speech technology can be especially seen in the area of foreign language education, which has led to the development of a new discipline known under the name of Computer-Assisted Language Learning (CALL). The literature on CALL mentions a number of its potential advantages: elimination of time limitations and dependence on the teacher, possibility to work at learners' own tempo and to store the user's profile to monitor the progress, access to a number of additional materials such as visualizations, recordings, animations, comments, and elimination of the stress related to the fact that the learner is being listened to by his/her classmates.

Due to lack of knowledge of adequate prosody processing both for linguistic and technology purposes, its inherent complexity and the ensuing difficulty in its acquisition, intonation and other prosodic phenomena like rhythm and voice quality were ignored in language teaching for many years.

There appear to be several reasons for the generally growing interest in intonation within last years. First, there have been important new advances in the theory of intonation, its functions and forms, aided by the growing accessibility of acoustic signal analysis, processing and interpretation. Second, the expansion of the analytical domains of traditional linguistics from sounds and words to larger units of inquiry such as phrases, discourses and interactions, has drawn attention to such subfields as pragmatics, discourse analysis, and conversation analysis. Finally, applied linguistics has grown to give priority to communicative function of prosody rather than its linguistic form.

The main goal of this paper is thus to integrate both segmental and suprasegmental aspects, especially in discourse and interaction, and to suggest a complex framework for studying foreign language prosody with Euronouce software.

2. Challenges in computer-assisted language learning

2.1. Audio and visual training

The advantages of multimedia tools in education are multiple. The use of tools for audiovisual feedback to detect deviations from standard articulation in the target language have shown especially high effectiveness in PC-based pronunciation learning systems. Prosody visualization seems to be more complex. However, speech analysis has been used for teaching L2 (second language) intonational patterns since 1970s for example [2] or [3]. The main principle is that the sound waveform or pitch contour of the student's utterance is visually displayed alongside those of the teacher, e.g. with Visi-Pitch by Kay Elemetrics, students are able to see both the model speaker's and their own intonational curve simultaneously.

The main technical shortcomings of hardware and software used currently for prosody training and research can be summarized as follows:

- Weak speech signals;
- No extrapolation for voiceless sounds;
- Not entirely correct/reliable F0 extraction;
- Lack of voice quality visualization.

Methodological shortcomings include:

- Lack of user-friendliness, i.e. learners do not know how to interpret displays and evaluate results;
- Examples and exercises consist of word and sentence-level intonation;
- Lack of integration of prosodic features as tone, duration, loudness;
- Lack of voice quality analysis - even if the learner can produce individual sound segments which are very similar to those produced by the teacher, they may still sound 'wrong' due to overall voice quality.

In order to develop an effective audio and visual training that improves learners' perception and production of intonation

and rhythm we need to better understand the relationship between perception and production at the level of segmentals and especially suprasegmentals. There are generally three types of prosodic phenomena which are not understood well enough to develop effective prosody training tools: 1) those which divide the speech into *chunks* or *units*, 2) those which lend *prominence* and 3) paralinguistic and nonlinguistic phenomena [4].

It should also be noted that the importance of auditory and/or visual feedback with regard to prosody is difficult to assess because computer programs providing feedback require from learners to be able to monitor and evaluate themselves critically. Apart from visual display, no further feedback is provided and there is a lack of objective assessment. Another question concerns the long-term effects of any of the brief training sessions.

2.2. Speech technology in language learning

Advanced PC-based learning systems (like Pronunciation Power, American Sounds, Phonics Tutor, Eyespeak) include (verify [5]): 1) speech analyzing windows or frames, 2) Internet-based features like email answering, online help and chat sessions with human tutors, 3) animation of the articulatory mechanics, video clips showing jaw, lip and tongue movements and waveform patterns of sound samples. Users can record sound files and acoustically compare a graphical representation of their utterances with those of the instructor. A few systems (e.g., Fonix iSpeak 3.0, Pro-Nunciation) include synthesized speech or TTS solutions [6].

During the last decade ASR technology was implemented into innovative interactive systems like Istra and Pronto [7] and in research projects such as ISLE [8]. The recognizers were trained and tested on non-native speech and researchers tried various probabilistic models to produce pronunciation scores from the phonetic alignments generated by HMM-based acoustic models. In few cases, like in [9], also prosodic features were taken into account. In the FLUENCY project [10] correlation between pronunciation and prosody errors was investigated. However, neither the placement of the intonation errors, nor suggestions on how to improve intonation were provided, leaving the comparison to the users.

Well-known software, Tell Me More of Auralog, improved the detection and feedback for pronunciation practice by pointing out erroneous phonemes and showing a 3D animation of the 'standard' articulation. However, its technology for suprasegmentals is very limited.

3. Towards optimized technology for L2 prosody training

3.1. Euronounce project

Intelligent Language Tutoring System with Multimodal Feedback Functions (acronym Euronounce) aims at creating L2 pronunciation and prosody teaching software. The project focuses on Slavonic (including Polish, Slovak, Russian and Czech)-German language pairs. The Euronounce project was preceded by earlier projects carried out by between 2005 and 2007. As a result, an audiovisual software AzAR (German acronym for *Automat for Accent Reduction*) aimed at teaching Russians German pronunciation was created [1]. Following the baseline developed in these projects the Euronounce project beside new language pairs adds also suprasegmental exercises.

3.2. Speech databases and text corpora

In the development of the system considerable multilingual speech databases were created containing native speech in all 5 languages for automatic speech recognizer as well as complex non-native speech databases including spontaneous speech, continuous speech as well as simple and complex sentences designed to investigate specifically selected phenomena. The most innovative part of the corpus is *prosodic test*. Its purpose is to investigate the realization of prosodic/intonational features and L1 interferences in the domain of prosody. The text material for the prosodic test was created according to the same criteria as the exercises for prosody training described in sec. 3.4. Prosody test was recorded both by non-native and native speakers. The resulting speech material serves as a reference for the assessment of non-native prosody.

3.3. Annotation of speech databases

The whole speech material was segmented and phonetically transcribed using force alignment. Part of the non-native speech database has been manually verified i.e., segment boundaries were adjusted, noises, disfluencies and pauses were marked, the transcription and automatically inserted primary and secondary stress markers were checked. Finally, deviations from the canonical pronunciation (insertions, deletions and substitutions) were marked.

3.4. Suprasegmental features

In accordance with the current emphasis on communicative and sociocultural competence, more attention should be paid to discourse-level communication and to cross-cultural differences in pitch. As natural discourse exhibits anything but "default" intonational patterns, L2 learners must be made aware of how stress, emphasis, contrast, and illocutionary speech are expressed in L2. In order to meet these goals, the system under development contains a module for prosody training.

The exercises are devised in order to test and practice prosody in smaller and larger syntactic units. In isolated words suprasegmental identification is devoted mainly to the perception and production of regular and irregular lexical stress and foot structure as well as types of nuclear accents, duration, intensity, identification of mono-, di-, tri-, four-syllable words, prosodic word, enclitics, proclitics, linking.

At the level of simple and complex sentences exercises consist in production and recognition of different types of sentences, i.e. declaratives, commands, wh-questions, etc. on the basis of their suprasegmental features. Also building the awareness of the relationship between focus and meaning needs close attention. Identification and production of emphatic stress, relating focus with meaning and performing communicative functions with focus should be practiced e.g. showing emotions, disagreement, calling attention to new information. Special attention is also given to contrastive pitch patterns conveying various meanings: fall (finality, authority), rise (unfinished, insinuating, tentative), level (unfinished, unresponsive), fall-rise (reservation, contrast, calling), rise-fall (insistence, surprise, irony).

4. Feedback system

4.1. Segmental structure

Lack of proper (or any) feedback is often named as the most serious flaw in educational software [7], [11]. Good software should not only assess the correctness of pronunciation but also instruct on how to improve it, show where exactly the error has been made, and offer feedback that is easy to interpret. To answer these needs the Euronounce software provides a multimodal feedback which includes visual and audio modules in the form of curriculum recordings by a reference voice and the visualization of the speech signal under the transcribed and phonemically segmented reference utterances. The software uses HMM-based speech recognition and speech signal analysis on the learner's input which makes a visual and aural comparison of the user's own performance with that of the reference voice possible. Most importantly, the system also performs an automatic error detection on the phonemic level. All uttered phones are marked using a color scale. Additional visual mode includes animated visualization of the vocal tract (lips area and articulators movements) and a formants graph for particular phones. A typical AzAR template for an exemplary phrase is shown in Fig. 1 and 2.

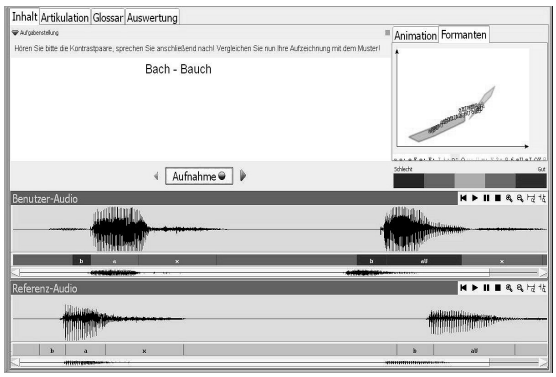

Figure 1: AzAR template for pronunciation assessment of a minimal pair Bach – Bauch (DE).

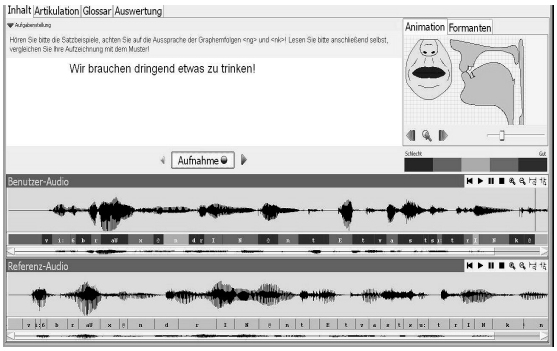

Figure 2: AzAR template for pronunciation assessment of an exemplary phrase wir brauchen dringend etwas zu trinken (DE).

Positive results of this kind of audio-visual feedback have been reported especially in prosody teaching [12], [13]. For pronunciation training [14] also traditional instruction is being recommended since visuals can be too difficult for the user to interpret and listening drill is not enough when one keeps in mind that L2 learner tends to associate foreign sounds with more familiar L1 sounds. Therefore, beside audio-visual feedback, AzAR software includes also text tutorial on articulatory and basic acoustic phonetics with glossary, phonemes description and classification, anatomic information, etc.

4.2. Suprasegmental structure

In order to provide an effective feedback on prosody, software should visualize the "relevant" intonation pattern of a given utterance as realized by L2 student and native speaker. It should also draw attention to acoustic features involved in the realization of intonation [15]. For example, the software could (a) instruct learners to compare the steepness of their falling or rising pitch movement to that of the native speaker, and/or (b) provide a quantitative measurement of the actual pitch slopes of both the native speaker and the learner. An effective feedback of this kind requires implementation of some kind of pitch stylization and normalization. Pitch Line program designed for approximation and parameterization of intonation contours answers these needs and could be successfully implemented in the AzAR environment.

The method behind the Pitch Line stylization is based on the assumption that intonational tunes can be regarded as *strings* of *events* (pitch accents, boundary tones) associated with the segmental structure of the utterance. The events are modeled as rising, falling or rising-falling pitch movements. They are delimited by target points in the contour (F0 minima and maxima) which define their start, peak and end; some of the targets are effectively corresponding to phonological tones (H, L). At the moment, identification of pitch targets' position is carried out manually. The parts of F0 contour corresponding to the events are approximated with functions described as follows:

$$\begin{aligned} 0 < x < 1 \qquad & y = x^{\gamma} \\ 1 < x < 2 \qquad & y = 2 - (2 - x)^{\gamma} \end{aligned} \tag{3}$$

The stretches of contour between subsequent events are called *connections* and are approximated with straight lines. In Pitch Line the approximation is carried out semi-automatically: the choice of the approximation function i.e., R-rising, F-falling, or C-connection (cf. [16]) and the alignment of the function with the segmental string depend on the human labeler and are decided upon by clicking in the appropriate location on the approximation panel. It is assumed that the start and end of the approximation functions have to be aligned with some segmental landmark located on the pre-accented, accented or post-accented syllable. During the approximation the normalized mean square error can be controlled: it is displayed on the approximation panel.

At the output the program provides a file containing the values of the stylized F0 curve (which can be used for pitch resynthesis in *Praat*) and another file with parameters describing the events: slope (describing the steepness of the F0 curve), Fp (F0 value at the point of the alignment of the approximation function), amplitude of the pitch movement and shape coefficient of the curve.

Fig. 3 illustrates the editing window of Pitch Line (from top to bottom): waveform, SAMPA transcription, the original

(dotted line) and stylized F0 contour (solid line), approximation functions (R,F, C) and NMSE. The vertical lines show approximate phoneme boundaries.

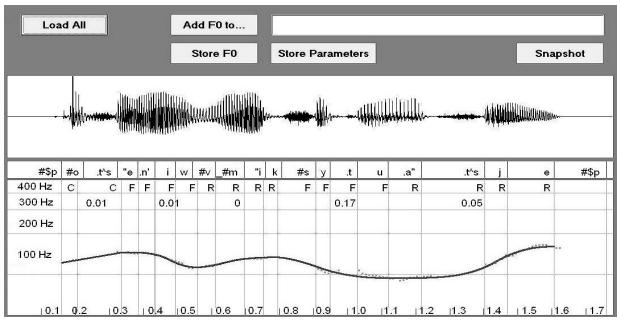

Figure 3: The editing window and stylization of an intonation contour with the Pitch Line program.

The usefulness of the approach adopted in Pitch Line was tested on a speech corpus including recordings of two speakers (male and female) reading a novel passage (1000 phrases), see [17]. The stylization accuracy was evaluated objectively by measuring the NMSE value between original and stylized F0 contours and subjectively in a perception study. The average NMSE value for the two speakers is 0.003, which indicates that the proposed method provides an accurate approximation of F0 contours. The general impression of the listeners was that the phrases resynthesized with the stylized F0 contours sounded very natural. It shows that Pitch Line stylization is well capable of extracting the macroprosodic component of F0 curves reflecting the choice of the intonation pattern for the utterance. Informal tests were also carried out on a subset of German utterances from the Euronounce speech database. The work is in progress to develop automatic pitch target detection so that fully automatic stylization is possible. If that succeeds, the implementation of Pitch Line in the AzAR training environment can be considered.

5. Conclusion

This paper presents general specifications and more detailed assumptions for suprasegmental training for the Euronounce project which aims at creating an intelligent system with multimodal feedback functions for learning pronunciation and prosody of German, Polish, Slovak, Czech and Russian as L2. The article outlines basic theoretical foundations as well as the core technology established in the preceding projects based on the German-Russian language pair. This baseline coupled with new cross-lingual databases are to help improve the visualization and quality assessment methods and to allow including prosodic factor in the final software.

6. Acknowledgements

This project has been funded with support from the European Commission within the Lifelong Learning Programme (project 135379-LLP-1-2007-1-DE-KA2-KA2MP). The project homepage is located at: http://www.euronounce.net.
The computer implementation of Pitch Line was done by J. Ogórkiewicz from the Laboratory of Language and Speech Technology within the framework of a national R&D project no. R00 035 02.

7. References

[1] Jokisch, O., Koloska, U., Hirschfeld, D. and Hoffmann, R. “Pronunciation learning and foreign accent reduction by an audiovisual feedback system”. *Proc. 1st Intern. Conf. on Affective Computing and Intelligent Interaction (ACII),* Beijing, China, 2005, pp. 419-425.

[2] Abberton, E. and Fourcin, A. J. , "Visual feedback and the acquisition of intonation", In E. H. Lenneberg & E. Lenneberg (Eds.), *Foundations of Language Development* (New York: Academic Press, 1975), pp. 157-165.

[3] de Bot, K. and Mailfert, K., "The teaching of intonation: Fundamental research and classroom applications", *TESOL Quarterly, 16*, 1982, 71-77.

[4] Cruttenden, A., *Intonation*, Cambridge University Press, Cambridge, 1997

[5] Learning Village. Educational Software Review, Retrieved on 15th July 2008 from http://www.learningvillage.com/html/guide.html

[6] Burston, J., The CALICO Software Review. Computer Assisted Language Instruction Consortium homepage. Retrieved on 15th July 2008 from http://calico.org/CALICO Review/

[7] Dalby, J. and Kewly-Port, D., "Explicit Pronunciation Training Using Automatic Speech Technology", *CALICO Journal, 16* (3):425-445, 1999

[8] Interactive Spoken Language Education (ISLE), project homepage. Retrieved on 15th July 2008 from http://nats-www.informatik.uni-hamburg.de/~isle/

[9] Teixeira, C. Franco, H., Shriberg, E., Precoda, K. and Sönmez, K. "Prosodic Features for Automatic Text-Independent Evaluation of Degree of Nativeness for Language Learners". *Proc. 6th ICSLP*, Beijing, China, 2000, pp. 187-190

[10] Eskenazi, M. and Hansma, S., "The Fluency pronunciation trainer", *Proc. Speech Technology in Language Learning*, Marholmen, 1998, pp.77-80.

[11] Engwall, O. Wik, P., Beskow, J. and Granström, G. "Design strategies for a virtual language tutor". *Proc. 8th ICSLP,* Jeju Island, 2004, pp. 1693-1696.

[12] Eskenazi, M. "Using automatic speech processing for foreign language pronunciation tutoring: some issues and a prototype", *Language Learning & Technology, 2*(2):62-76, 1999.

[13] Chun, D..M., "Signal analysis software for teaching discourse intonation", *Language Learning & Technology*, *2*(1):61-77, 1998.

[14] Neri, A., Cucchiarini, C. and Strick, H." Feedback in Computer Assisted Pronunciation Training: When technology meets pedagogy". *Proc. 10th Int. CALL Conference on "CALL professionals and the future of CALL research",* Antwerp, 2002, pp. 179-188.

[15] t'Hart, J., Collier, R. and Cohen, A., *A Perceptual Study of Intonation,* Cambridge University Press, 1990

[16] Taylor, P. "Analysis and synthesis of intonation using the tilt model", *J. Acoust. Soc. Am 107*(3):1697-1714, 2000.

[17] Wagner, A. "A comprehensive model of intonation for application in speech synthesis", *Proc. 8th International PhD Workshop OWD*, Wisła, Poland, 2006, pp. 91-96